Stepping into the Boardroom: A novel AI-enabled framework for recognising empirical manifestations of Collective Leadership from textual data

Anonymous ACL submission

Abstract

The concept of Collective Leadership (CL, broadly speaking leadership within groups) is difficult to define and detect empirically. A promising avenue for detecting CL focuses on discursive approaches based on group interaction and 'turning points' in the discussion, where participants concur on the need for action. In the absence of a defined NLP task for the detection of CL, we present a novel AI-enabled pipeline applied to publicly available hospital board text data, requiring minimal annotation thanks to in-context learning. To our knowledge, this research is the first to combine NLP and leadership theories. After presenting a language model architecture, we propose an experimental approach using ablation analysis and posit an evaluation set-up including a 'human in the loop'to aid acceptability by organisational research scholars and support the development of an annotated dataset.

1 Introduction and related work

The literature surrounding Collective Leadership includes ample theorising but limited research on how it manifests empirically, let alone in the context of executive boards (Edwards and Bolden, 2023; Croft et al., 2022; Ospina et al., 2020; Fairhurst et al., 2020).

Croft et al. (2022) define CL as:

```
"The interaction of strategic ambiguity
and inward- and outward-facing
reification practices to maintain
divergent perspectives alongside agreed
collective aims, alignment, coordination
of activities, and commitment
to collective success."
```

A promising avenue for detecting CL and connected concepts in the above definition (such as strategic ambiguity, reification and collective work) includes discursive approaches to leadership, interaction and 'turning points' in a discussion, where participants concur on the need for action (Fairhurst, 2007; Sklaveniti, 2020; Lortie et al., 2022). These discursive approaches have yet to make use of Natural Language Processing (NLP) techniques to detect CL. Our review of the NLP literature on group decision-making Mayfield and Black (2019b,a, 2020) identified only one dataset, the Wikipedia's Article for Deletion forums (Xiao and Sitaula, 2018; Xiao and Nickerson), and no definition of an NLP task specific for detecting CL. Overall, these findings reflect that NLP (or large-scale text analytics) is hardly applied in the domain of organisational research or leadership studies (Hannigan et al., 2019). Against this background, in this study we seek to respond to this research question: *"In the absence of a defined NLP task for the detection of CL, what is the most appropriate, AI-enabled pipeline for identifying CL using solely board meeting textual data (board reports, minutes)?"*

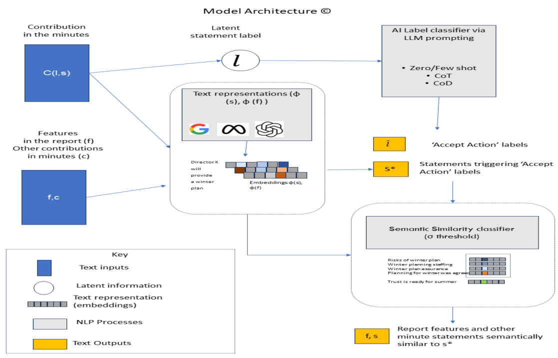

Figure 1: Proposed NLP architecture. The model receives text from the minutes and reports to create embeddings. In the first instance, a text-classification tool processes the minutes to identify formal actions ('Accept Action' label). Sections of the minutes classified with that label go through a semantic similarity classifier to identify other similar texts in future board reports/minutes.

2 Methods: A novel NLP task to identify CL

2.1 Preliminaries: a conceptual methodology for identifying CL in board text data

To inform our NLP approach, we translate the concept of CL into an executive board space of NHS hospitals by focusing on particular 'turning points' in the discussion, where participants formally agree on the need for change through a minuted action. From an NLP perspective, we initially seek to identify or classify sections of minutes which the language model can label as 'Accept Action': not only an action has been requested, but it has been formally allocated to an individual and recorded as such in the minutes. We do this based on an adaptation of the ISO standard for dialogue act annotation (ISO 24617-2:2012).

Following Croft et al.'s (2022) definition above, we posit that to detect collective leadership from executive board text requires the fulfilment of two conditions:

1. **Collective work** (joint understanding over time across teams) (Gronn, 2000): This means members of the board actively discuss an issue already highlighted in a report for that meeting, and references to an issue are seen over time (i.e. across several board sessions/reports, both in past and future). This signals a level of sustained, synergistic understanding and coordination between managers, executives and non-executives within NHS Boards.

2. **Evidence of reification over strategic ambiguity** (commitment and focus towards an aim): As noted above only those discussion points where there is reification in the form of a formally recorded action (a latent 'Accept Action' label) can signal CL, as these noted actions formally task managers and executives to prioritise their activities over other conflicting demands Croft et al. (2022). Consistent with the point above on collective work, we add there must be a follow-through on that specific action in future.

These degrees of reification and the distinction between collective work and collective leadership are illustrated in Figure 2. We translate these degrees of reification into a hierarchical taxonomy or ranking of 'discussion labels'. Our analysis focuses on the discussion label 'Accept Action' - a minuted action allocated to an individual. This label is more formally introduced in the next section to formulate CL mathematically, drawing from Mayfield and Black's (2019a) notation.

2.2 Proposed approach

From the point of view of the meeting space, we can identify CL when we see an 'Accept Action' label as part of a discussion in the minutes, provided that features of that discussion will have some follow-up over time (in future), and there has been some discussion about it (contemporaneously or in the past). This is formalised in equation 2.1 below.

$$\begin{aligned}\exists \tau_1, \tau_2, \in \{1, \ldots, \tau_1, \ldots, t \ldots, \tau_2, \ldots, T\},\\ l_i^t = \text{'Accept Action'}\\ \text{and } s_i^{\tau_1} \sim f_i^{\tau_1} \sim s_i^t \sim f_i^t\\ \text{and } s_i^t \sim f_i^t \sim s_i^{\tau_2} \sim f_i^{\tau_2}\end{aligned} \tag{2.1}$$

Applying a label 'Accept Action' at time t for a paragraph or section of the minutes s_i^t in isolation does not reflect collective leadership; it does so only if (i) we see other sections of minutes or sections of reports with a similar topic in future, ($s_i^{\tau_2}$,$f_i^{\tau_2}$), sustained commitment over time through discussion/follow-up and (ii) it has also been raised previously in minutes or reports ($f_i^{t<\tau_1}$) - reflective of our 'collective work' condition above. Equation 2.2 shows there must be at least two points in time (in past -τ_1 and in future -τ_2) where the language model is able to identify semantic similarity compared to the text (s_i^t) in time t which as been classified with an 'Accept Action' label.

2.3 Task definition

The **input** is the dataset containing our corpus of board-level documents (reports and minutes, split in paragraphs f and s respectively) for each hospital h for the period 2017-2023. We also require a set value for the parameter sigma. In our notation σ is a parameter that denotes a quantitative threshold for similarity (such as Dice Coefficient or Jaccard Index (Peinelt, 2021)). As part of our experimental setting, we will test various levels of σ. Below we identify each report and minutes.

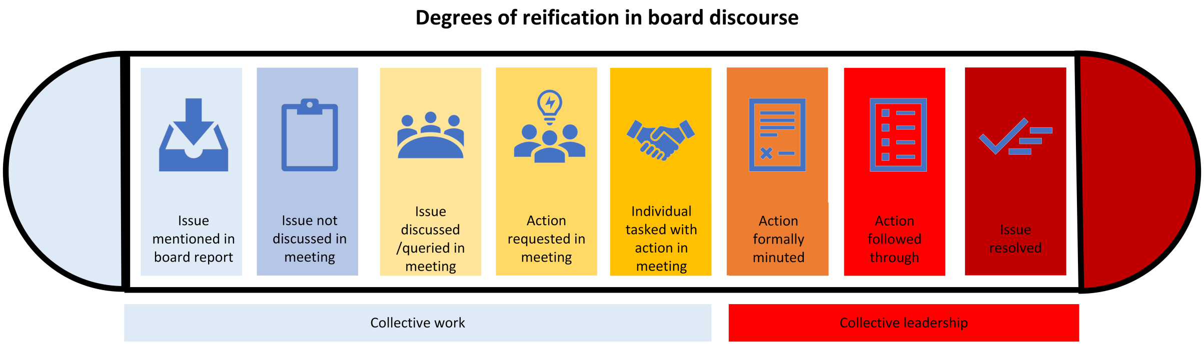

Figure 2: Degrees of reification in board discourse. The chart outlines the distinction between collective work and collective leadership

$$
\begin{aligned}
R_{ih}^{t} &= (l, f_1, f_2, \ldots, f_F)_{ih}^{t} \\
M_{ih}^{t} &= (c_1, c_2, \ldots, c_C)_{ih}^{t} = \\
&= ((\mathfrak{l}, s_1), (l_2, s_2), \ldots, (l_c, s_c))_{ih}^{t}
\end{aligned}
\tag{2.2}
$$

We defined above the text tuples R and C. These tuples are composed of a latent (unknown) action label l and a set of paragraphs (which we call features f - in reports or statements s - in the minutes).

The **objective** of this task is to identify CL as defined in equation 2.1 above for a particular t in $\mathcal{T}$, h in $\mathcal{H}$. We do this by first identifying reification over strategic ambiguity:

$$
\begin{aligned}
&\widehat{l_{ih}}^{t} = \mathbb{E}({l_{ih}}^{t} = \text{'Accept Action'} | \phi({s_{ih}}^{t})) \\
&\text{for any } c_{ih}^{t} = (l_i, s_i)_h^t \\
&s_i^* = s_i^t \quad \text{where} \\
&\widehat{l_i}^{t} = \mathbb{E}({l_i}^{t} = \text{'Accept Action'} | \phi({s_i}^{t}))
\end{aligned}
$$

(2.3)

Once we have identified the relevant section of the minutes dealing with a particular action, we verify the condition of collective work over time. We do this by identifying that a similar text which can be found in other minutes and reports at other points in time (in future).

$$
\begin{aligned}
\sigma_{ij} &= \operatorname{Sim}\left(\phi\left(s_i^*, s_j^{\tau}\right)\right) > \sigma \quad \text{and} \\
\varsigma_{ij} &= \operatorname{Sim}(\phi(s_i^*, f_j^{\tau})) > \sigma
\end{aligned}
\tag{2.4}
$$

For at least one j and two τ: one $\tau_1 < t$ and $\tau_2 > t$, as per equation 2.1.

Algorithm 1 Summary procedure for Identifying Collective Leadership from a Single Set of Reports for Hospital h, Time t

---

Input: Corpus $= (R_{ih}^t, M_{ih}^t)$ for all reports i in I and all t in T; threshold σ; label set L.
Output: $O_h^{\tau} = \{s_i^*, s_j, f_j\}_h^{\tau}$

1: **return** s_* ▷ Identified relevant sections with 'Accept Action' label
2: s_j.append(s_j) for $\tau_1 < t$ ▷ Identified relevant minutes sections where discussions took place
3: s_j.append(s_j) for $\tau_2 > t$ ▷ Identified relevant sections in future meetings where the action is followed up
4: f_j.append(f_j) for $\tau_1 < t$ and $\tau_2 > t$ ▷ Identified relevant paragraphs in the minutes and reports with similar semantic similarity to each element of $s_\star$
5: **return** O_τ

2.4 Model architecture and experimental design

Figure 1 outlines the proposed architecture of the language model. Within the architecture, we have considered various potential text representations and prompting approaches as part of our experimental design.

Our experimental design considers 12 (3x4x3) architectures as outlined below:

- **Training Dataset:** We will train the classification model using in-context learning through a small manually labelled dataset drawn from minutes from a single NHS hospital not included in our sample, splitting the dataset in 80\10\10 proportion. We will aim to have at least 10 examples of each label as per Brown et al. (Brown et al., 2020), while aware the source data is heavily skewed towards 'discussion' and 'query' labels instead of 'action' labels.

- **Text representations:** We will consider three different text representations using GPT-4 (by Open AI)(OpenAI, 2023), LlaMa 2 (by Meta)(noa) and BERT (by Google) (Devlin et al., 2019).

- **Prompts:** We will consider four different prompting methods: zero/few-shot (Brownlee, 2018), chain of thought prompting (Wei et al., 2023), chain of density (Adams et al., 2023).

- **Semantic Similarity**. As this is a standard NLP task, we propose to use a single architecture, tBERT (Peinelt, 2021), but testing 3 thresholds for similarity.

2.5 Evaluation

When developing a CL-detection NLP task we face an evaluation challenge as there is no established 'ground truth', something to benchmark the model against. In this case, the NLP literature suggests a combination of quantitative, qualitative and human-based evaluation techniques (Mayfield and Black, 2019a), which we apply to the various components of the task as well as the task overall.

Below we propose our evaluation framework, with the evaluation following in subsequent work.

- Quantitative (multi-class classification): Following (Brown et al., 2020), for zero/few-shot learning, we evaluate the consistency of the classification through random five-fold cross-validation against the small training dataset we have developed.

  Our main evaluation metric is Balanced Accuracy, which is more sensitive to smaller class sizes. This is helpful as we have seen from preliminary review of data that the 'Accept Action' labels are less frequent than others. As a complementary metric we will consider the specific F1 for 'Accept Action' as a secondary metric, implicitly simplifying the classification problem from multi-class to binary.

- Quantitative (Semantic Similarity): We will use the standard F1 metric, but mindful of the lexical overlap bias found in semantic similarity tasks we consider the 'non-obvious F1' metric introduced by Peinelt (Peinelt, 2021), as a complementary metric. We will evaluate the consistency of the classification through random five-fold cross-validation throughout the small dataset constructed against Spacy's dependent parser model which has been found to have satisfactory performance in unsupervised settings.

- Qualitative human validation: Consistent with the Computational Grounded Theory method (CGT), selected passages identified as CL will be subject to human deep-reading to validate the findings.

3 Conclusion

In this work we have established a text-based definition of collective leadership to motivate our proposed novel AI-enabled NLP task, including algorithms, a proposed database, architecture, experimental design and evaluation for detecting CL from board text data.

4 Limitations and broader impact

In terms of risks, this research was not subject to ethics approval given that the source data is in the public domain. However, in terms of broader impact, we still have duties of confidentiality and are mindful of potential professional implications for the NHS executives and officials who are part of these boards, given the potential implications of our analysis of CL on ongoing transformation activities.

The abstract and introduction to our paper summarise our main claims regarding the lack of an pre-established NLP task for the detection of CL. While we have based our approach on other NLP literature on group decision-making Mayfield and Black (2019b,a, 2020), there are limitations on greenfield cross-disciplinary research on organisational research concepts such as CL. This includes the lack of consensus on a definition of CL, and the historically qualitative methodologies used to evaluate it. We have proposed an experimental and evaluation approach including ablation analysis that keeps a human in the loop. This qualitative element is meant to: (i) appeal to organisational researchers less familiar with NLP techniques, (ii) support validation as there is no established 'ground truth' or benchmark, and, (iii) aid the development of labelled datasets for the NLP problem of CL detection.

We also explore potential limitations arising from various biases and challenges with reproducibility and accuracy arising from these methods.

1. **Methodological biases and errors** might emerge through the pre-processing (encoding) of the textual data. We seek to minimise these biases by undertaking different approaches to encoding the textual data and establishing clear evaluation metrics for each. Research has also identified that Large Language Models might be biased towards outputs that mimic frequent training examples (Jones and Steinhardt). We sought to minimise this by providing a balanced set of decision-making label training examples.

2. **Data biases.** Minutes, committee documents and routine reports are classified as "reportative" (Heller, 2023) sources containing factual, historical information, but we recognise limitations related to 'authorship, bias and power' used in those documents (Heller, 2023). Large Language Models, as repositories of language data, include social biases around gender, race, religion and social constructs (Liang et al., 2021).

3. **Researcher bias**. Critical to any research design is that it intimately reflects the researcher's perspective, which is shaped by their own beliefs and the scientific community they belong to (Kaur and Kumar, 2021). In agreement with CGT, we mitigated this by approaching the AI analysis iteratively and in a phased manner. CGT can help avoid biased interpretations of qualitative data because of this iterative approach back and forth between the human analyst and the computational analysis, instead of the failed presumption that quantitative approaches are bias-free (particularly given the use of natural language to 'prompt' the AI) (Tschisgale et al., 2023).

4. **Reproducibility.** A limitation of the approach is reproducibility, while most of the computational steps are reproducible through access to the software, there is an interpretation in the qualitative coding that supports grounded theory. In establishing CGT, Nelson (Nelson, 2020) recognises that faced with the same computationally enabled results, the researcher might not code these in the same way.

5. **Accuracy of pre-trained language models.** Our approach intends to build upon pre-trained large language models which are domain-agnostic. While pre-trained models using domain-specific, our limited pre-annotated data might not be able to achieve higher levels of accuracy and performance, given the large cost in serving the 'long tail' of other domains (Tschisgale et al., 2023). Our training is limited to the labelling of a small section of out-of-sample board reports as to achieve a handful of examples of the different types of 'discussion labels' to classify sections of the minutes.

Acknowledgements

Thank you in advance to anonymous ARR reviewers and to [Anonymous] collaborators for their feedback.

References

Llama 2: Open Foundation and Fine-Tuned Chat Models | Research - AI at Meta. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

Griffin Adams, Alexander Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Jason Brownlee. 2018. A Gentle Introduction to k-fold Cross-Validation.

Charlotte Croft, Gerry McGivern, Graeme Currie, Andy Lockett, and Dimitrios Spyridonidis. 2022. Unified Divergence and the Development of Collective Leadership. *Journal of Management Studies*, 59(2):460–488.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Gareth Edwards and Richard Bolden. 2023. Why is collective leadership so elusive? *Leadership*, 19(2):167–182.

Gail T. Fairhurst. 2007. *Discursive Leadership: In Conversation with Leadership Psychology*. SAGE Publications, Incorporated, Thousand Oaks, UNITED STATES.

Gail T Fairhurst, Brad Jackson, Erica Gabrielle Foldy, and Sonia M Ospina. 2020. Studying collective leadership: The road ahead. *Human Relations*, 73(4):598–614.

Peter Gronn. 2000. Distributed Properties: A New Architecture for Leadership. *Educational Management & Administration*, 28(3):317–338.

Timothy R. Hannigan, Richard F. J. Haans, Keyvan Vakili, Hovig Tchalian, Vern L. Glaser, Milo Shaoqing Wang, Sarah Kaplan, and P. Devereaux Jennings. 2019. Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*, 13(2):586–632.

Michael Heller. 2023. Rethinking Historical Methods in Organization Studies: Organizational Source Criticism. *Organization Studies*, 44(6):987–1002.

Erik Jones and Jacob Steinhardt. Capturing Failures of Large Language Models via Human Cognitive Biases.

Manmeet Kaur and Rajesh Kumar. 2021. Mixed Methods in Global Health Research. In *Handbook of Global Health*, pages 239–260. Springer, Cham.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6565–6576. PMLR.

Josée Lortie, Laure Cabantous, and Cyrille Sardais. 2022. How Leadership Moments are Enacted within a Strict Hierarchy: The case of kitchen brigades in haute cuisine restaurants. *Organization Studies*, page 01708406221134225.

Elijah Mayfield and Alan Black. 2019a. Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 65–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Elijah Mayfield and Alan W. Black. 2019b. Analyzing Wikipedia Deletion Debates with a Group Decision-Making Forecast Model. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

Elijah Mayfield and Alan W Black. 2020. Should You Fine-Tune BERT for Automated Essay Scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Laura K. Nelson. 2020. Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, 49(1):3–42.

OpenAI. 2023. GPT-4 Technical Report.

Sonia M Ospina, Erica Gabrielle Foldy, Gail T Fairhurst, and Brad Jackson. 2020. Collective dimensions of leadership: Connecting theory and method. *Human Relations*.

Nicole Peinelt. 2021. *Detecting Semantic Similarity : Biases, Evaluation and Models*. Ph.D. thesis, University of Warwick.

Chrysavgi Sklaveniti. 2020. Moments that connect: Turning points and the becoming of leadership. *Human Relations*, 73(4):544–571.

Paul Tschisgale, Peter Wulff, and Marcus Kubsch. 2023. Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research*, 19(2):020123.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Lu Xiao and Jeffrey V Nickerson. Imperatives in Past Online Discussions: Another Helpful Source for Community Newcomers?

Lu Xiao and Niraj Sitaula. 2018. Sentiments in Wikipedia Articles for Deletion Discussions. In *Transforming Digital Worlds*, Lecture Notes in Computer Science, pages 81–86, Cham. Springer International Publishing.

A Use of AI assistants

AI assistants were used to support LATEX coding of equations and algorithms.